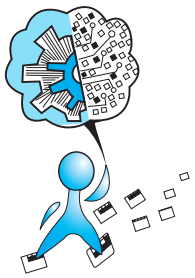


Data Pipelines architecture and basic understanding

Lohit Ravi Teja Bhupati

Data Engineer, Walmart. Bentonville, AR. Email: lohit.raviteja008@gmail.com

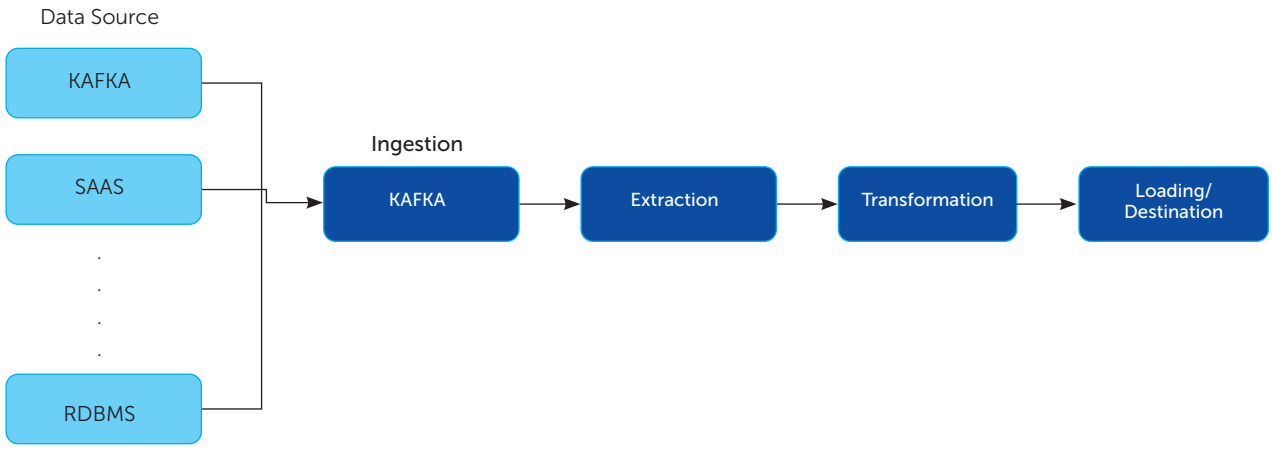
Data pipelines are the series of steps performed to process the data into required formats/ destination sources. In these pipelines the output of each step will be source to the following steps. There are different scenarios where the pipelines are designed, it can be a simple batch job to read raw data and convert it into formatted data and saving into sources tables, consuming data from the upstream applications and perform transformations so that the data could be utilized by the downstream applications or creating visualizations using the transformed data etc.

There are three major factors that contribute to the speed with which the data flows in a pipeline, **Throughput** can be defined as the rate at which the data flows through the pipeline. **Latency** is defined as the time required for a single unit/token of data to flow through the pipeline.

Reliability is the measure completeness and accuracy. There are different techniques that are in place to ensure the data reliability like logging, auditing, validating etc.

There are some basic questions one should have before building a pipeline.

- What data source?
- What type of ingestion?
- How frequent should the load happen?
- How should the end data look like?
- Is the data consumed by any downstream applications or any visualizations are to be built on top of it?



Data Pipeline architecture

Data Source:

It can be considered as the first and key layer of the design. It may include the data from any streaming sources, SaaS applications and relational databases etc.

Ingestion:

It is defined as a process to read the data from the data sources. We can read the data from each source using the API's provided by the data sources itself. Before reading the data, we need to analyze the data sets to gain insight

Contd. on page 18...

shown to help Zepto with. Without dark stores, last-mile connection would not have been possible due to the breadth of the country and the presence of people practically everywhere.

Managing the amount of supplies is another difficulty that India is accustomed to. Dark stores are normally used to stock a smaller range of goods, so Zepto must have given careful consideration to how to handle the variety that comes along with the volume of deliveries in each Indian metropolis.

Overall, Zepto has been successful with its setup and infrastructure approach and is on a solid growth trajectory.

Conclusion:

Zepto is making news because of how quickly it moves. Nevertheless, quick delivery does not sacrifice quality. One of India's start-ups with the quickest growth is this one. Customers are responding positively to the company's miraculous 10-minute product delivery. Their mission is to 'make 10-minute delivery normal'. Even though it seems so absurd to us, it's actually working...

About the Author

Saikumar Mediboina is currently pursuing his masters in Information Technology at the National Institute of Technology Karnataka Surathkal. His area of interest is networking. He is currently working in tactile internet to achieve the low end to end latency and high reliability Haptic media type with http and other protocols Synchronization of multi modal data distributed to multiple devices and location.

...Contd. from page 11

of the data through a process called Data Profiling, which is used to examine the quality and structure. The data can be ingested either batch ingestion or streaming ingestion. Batch Ingestion is a sequential processing where the set of data records can be extracted and processed together. These batch processes can be scheduled or triggered manually. In Stream ingestion a single record of data can be processed automatically from the data source as soon as it is created or in time windows which can be used to give near real time data.

This ingestion or extraction frequency depends on the requirement on how the subsequent application need the data to be loaded.

Transformation:

Once the data extraction is done, we can transform the data into the required format of the destination system, which helps to analyze the data.

Destination:

A destination is a data warehouse, a database or a data mart to hold the data after the data is processed.

About the Author

Lohit Ravi Teja Bhupati is currently working as Data Engineer at Walmart, Arkansas. He has done Masters thesis in Machine learning and Neuro imaging from University of Houston Clear lake, USA. With his interest towards Machine learning and data analysis, he is always upscaling by learning new technologies and continuing his passion.